



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



Running field experiments using Facebook split test

Davide C. Orazi^{a,*}, Allen C. Johnston^b

^a Dept. of Marketing, Monash University, Australia

^b Dept. of Info System, Statistics, & Management, University of Alabama, USA

ARTICLE INFO

Keywords:

Experimental research
Field study
Facebook
Split testing
Online advertising
Ecological validity

ABSTRACT

Business researchers use experimental methods extensively due to their high internal validity. However, controlled laboratory and crowdsourcing settings often introduce issues of artificiality, data contamination, and low managerial relevance of the dependent variables. Field experiments can overcome these issues but are traditionally time- and resource-consuming. This primer presents an alternative experimental setting to conduct online field experiments in a time- and cost-effective way. It does so by introducing the Facebook A/B split test functionality, which allows for random assignment of manipulated variables embedded in ecologically-valid stimuli. We compare and contrast this method against laboratory settings and Amazon Mechanical Turk in terms of design flexibility, managerial relevance, data quality control, and sample representativeness. We then provide an empirical demonstration of how to set up, pre-test, run, and analyze FBST experiments.

1. Introduction

Business research is witnessing an increasing tension between internal validity—the confidence that the observed relationships are causal (Campbell, 1957), and external validity—the confidence that results can be generalized to different populations and settings (Bell et al., 2018; Bracht & Glass, 1968; Inman, Campbell, Kirmani, & Price, 2018; Schram, 2005). While laboratory experiments using student samples and online experiments using Amazon Mechanical Turk (Daly & Natarajan, 2015; Paolacci, Chandler, & Ipeoritis, 2010) possess internal validity, their external validity is often limited by the artificiality of their settings (Schram, 2005) and, for online experiments, mounting issues of data quality (Ford, 2017; New Scientist, 2018). Field experiments represent a viable way to overcome issues of setting artificiality and to show that the focal effects persist in the real world (Inman et al., 2018). Yet field experiments have issues of their own: they are traditionally time- and resource-consuming and often lack internal validity due to environmental confounds that cannot be controlled for.

This paper aims to address the tension between internal and external validity in business research by introducing a cost-effective experimental setting to run online field experiments through the Facebook A/B split testing functionality (heretofore referred as “FBST”). The latest version of the FBST platform now features random assignment and overcomes previous issues related to audience overlaps and unequal distributions of age and gender across cells (for a critique, see Eckles, Gordon, & Johnson, 2018). The FBST thus possesses unique

methodological and practical strengths for experimental researchers interested in understanding which of two or more experimental conditions has the strongest effect on one or more managerially relevant dependent variables, such as click-through ratios, page engagements, and online purchases (Facebook, 2019). Through split testing, theoretical constructs of interest can be operationalized and manipulated either as images, texts, or videos, holding all other conditions constant. This feature renders the FBST an experimental setting attractive not only to advertising researchers, but to any business researcher studying constructs that can be operationalized through visual and audio-visual stimuli.

This primer is organized as follow. We first provide an overview of the FBST, detailing common misconceptions associated with Facebook experiments that the new split testing functionality helps to overcome. Next, we compare the FBST to laboratory and Amazon Mechanical Turk experimental settings in terms of design flexibility, ecological validity, sample representativeness, and data quality control. We then provide a step-by-step tutorial on how to set up a simple A/B split design through Facebook split testing functionality, how to pre-test the stimuli, and how to analyze the aggregate output through non-parametric tests. Next, we provide two empirical demonstrations for simple factorial designs and 2×2 designs with one manipulated and one measured factor. We conclude by discussing applications beyond advertising research, implications for business and information security research, and avenues for further methodological development.

* Corresponding author at: Department of Marketing, Monash University, 21 Sir Monash Street, VIC 3145, Australia.

E-mail address: davide.orazi@monash.edu (D.C. Orazi).

Table 1
FBST compared to laboratory settings and Amazon Mechanical Turk.

		Laboratory	MTurk	FBST
Design flexibility	Range of manipulated independent variables	High	Moderate	Moderate
	Range of dependent variables	High	Moderate	Low to Moderate
Managerial relevance	Ecological validity of stimuli employed	Low to High	Low to High	Moderate to High
	Ecological validity of delivery platform	Low	Low to Moderate	Moderate to High
	Ecological validity of dependent variables	Low to High	Low	High
Data quality control	Risk of experimenter effects	Low	None	None
	Risk of multiple responses by the same users	None	Moderate	Very Low
	Risk of speeding and cheating	Moderate	High	Very Low
	Risk of automated bot contamination	None	Moderate	Very Low
Sample Representativeness	Susceptibility to coverage error	High	Low	Very Low
	Risk of sample heterogeneity across fields	Moderate	Low	Low

Note: The assessment of defining characteristics of laboratory and MTurk samples combines Wade & Tingling's (2005) and Paolacci et al.'s (2010) comparative tables, and updates them with risks of speeding, cheating, and automated bot contamination as indicated by Ford (2017). The assessment of the defining characteristics of the FBST is based on the averaged evaluations of seven top experimentalists in the fields of business and information security research. Details on the expert evaluation procedure are available in *Web Appendix B*.

2. Facebook A/B split testing: an overview

The FBST, launched by Facebook in November 2017, allows advertisers to pre-test their online campaigns to optimize future advertising expenditures (Facebook, 2019). After randomly assigning two or more ads to a target population, advertisers can select the ad with the highest click-through rate, the lowest cost per click, or the highest number of page visits, depending on the objective they wish to achieve. Since the FBST is premised on random assignment, it is also advantageous to researchers interested in comparing two or more experimental conditions and understanding their effect on managerially relevant dependent variables. The range of experiments that can be conducted through the FBST is by no means constrained to advertisement testing only: any theoretical construct of interest that can be manipulated as images, texts, or videos can be embedded in the platform.

Historically, scholars have been cautious and even critical about experimental designs delivered through Facebook for two reasons: (i) confounds introduced by lack of random assignment in designs conducted before November 2017 (Eckles et al., 2018), and (ii) mechanisms producing endogenous variation at the user, targeting, and competition levels in experimental designs measuring overall campaign effectiveness (Gordon, Zettelmeyer, Bhargava, & Chapsky, 2018). Neither of these concerns are applicable to the platform presented in this paper, as discussed next.

First, designs predating November 2017 have been criticized because the lack of random assignment rendered ad delivery more dependent on delivery optimization algorithms (see Eckles et al., 2018, for a recent critique). Critics of this approach argue that several studies displayed unequal distributions in sociodemographic variables including age, gender, and education levels (Eckles et al., 2018). However, the sample size used in Facebook experiments tends to naturally inflate the significance of even the most trivial sociodemographic differences. For example, in their Study 1, Matz, Kosinski, Nave, and Stillwell (2017) report they reached 3,129,993 users with their campaign, and 10,346 users clicked the ad displayed. The study was later criticized by Eckles et al. (2018) for introducing unintended variance in their research design, but the significant differences in age distributions across conditions ranged between 0.3% and 0.7% — a trivial difference indeed, whose statistical power was amplified by a sample size of over 3 million users. We believe this difference does not undermine the validity of Matz et al. (2017) pioneering work. Regardless of our opinion, the introduction of split testing and a random assignment component largely eliminates the influence that optimization algorithms have on the delivery of test ads, making the proposed split testing approach a robust way to run experimental designs in a naturalistic, online field setting.

A second concern stems from Facebook experiments comparing a

treatment shown to the target audience, to a control group, which never sees the treatment, to estimate the overall campaign effect. This approach introduces systematic differences between treatment and control groups due to activity bias (Lewis, Rao, & Reiley, 2011), targeting optimization (Eckles et al., 2018), and competition-induced confounds (Gordon et al., 2018). All three biases are extensively discussed by Gordon et al. (2018) and concern the fact that, because the control group never sees the ad, individual levels of online activity, delivery optimization algorithms that show the campaign to audiences “more likely to fulfil the campaign’s objective,” (Eckles et al., 2018: 5245) and ads from “competitors” bidding for the same ad space can inflate the effectiveness of the treatment condition. None of these concerns applies to simple split testing as two or more ads are always shown to the target audiences. Any individual-, targeting-, or competition-induced confound is partitioned out through random assignment. We now turn on the advantages and disadvantages that the FBST possesses in comparison to other experimental recruiting methods.

3. Comparing the FBST to other experimental recruiting methods

In the following sections, we outline the strengths and weaknesses of experiments conducted through the FBST in terms of (i) design flexibility, (ii) managerial relevance, (iii) data quality control, and (iv) sample representativeness. In particular, we compare and contrast the advantages of the FBST against those of traditional laboratory experiments, and online experiments conducted on Amazon Mechanical Turk. Table 1 extends Paolacci et al. (2010) comparison between lab and Amazon Mechanical Turk experiments with an assessment¹ of the relative standing of FBST experiments.

3.1. Design flexibility

The main limitation of the FBST methodology comes from design flexibility. The field experiment nature of the FBST limits the range of independent variables that can be manipulated in comparison to traditional and MTurk experiments. Any theoretical construct that can be manipulated through text, pictures, or audio-visual material can be embedded in a FBST experiment. However, more complex methods to manipulate independent variables (e.g., writing tasks, scenarios, sentence scrambling, tasks that require high users’ involvement) cannot be implemented in the FBST and are best suited for laboratory and online

¹ Rather than rely on our own subjective elaboration, the assessment of the strengths and weakness of the FBST relative to other experimental settings was operated by interviewing ten top marketing and information security scholars well-versed in experimental design. The interview procedure is detailed in *Web Appendix B*.

experiments. The greatest limitation of the FBST comes from a comparatively restrictive set of dependent variables provided by the platform (FBST dependent variables are discussed in detail in [Section 4. Analyzing FBST Data](#)). In contrast, laboratory experiments allow for a greater range of dependent variables, including self-reported measures, product choices, reaction times, biometric outputs, and many more. MTurk experiments are more limited than laboratory experiments in terms of range of dependent variables; biometric outputs such as heart-rate variability metrics are hardly obtainable in an online setting. However, MTurk experiments allow the use of many proxy variables for real behaviors (e.g., purchase intentions), resulting in a larger set of potential dependent variables than the FBST. In summary, the FBST suffers from design restrictions typical of field experiments and constrains design options to the delivery platform employed (i.e., Facebook). However, it is this “forced” ecological validity that makes the FBST a managerially relevant tool that can fruitfully complement, rather than substitute, laboratory and online experiments, as we discuss next.

3.2. Managerial relevance

The key strengths of the FBST comes from the high ecological validity it provides in terms of (i) the stimuli employed, (ii) the delivery platform, and (iii) the dependent variables. In terms of the stimuli employed, the FBST has ecological validity comparable to laboratory and MTurk experiments that use real-world stimuli. Real-world stimuli in the form of advertisements can be employed in all three methods, with the FBST “forcing” this choice due to the nature of the delivery platform. A second strength comes from an ecologically-valid delivery platform, which happens on a social media platform with 1.66 billion daily active users as for January 2020 ([Zephoria, 2019](#)). Because FBST experiments occur in a natural online setting, they possess superior ecological validity compared to both laboratory and MTurk experiments. Laboratory settings in particular have been criticized for issues of setting artificiality ([Schram, 2005](#)). Lastly, the FBST uses a set of limited, yet ecologically valid, dependent variables including, among others, click-through, page engagements, and online subscriptions. With digital advertising expenditures surpassing television in 2017 ([Gordon et al., 2018](#)), the performance metrics FBST records as output represent managerially-relevant dependent variables widely used by business researchers and indexical of message effectiveness and motivation (i.e., click-through and page engagements) or reflective of actual behavior (e.g., online subscriptions). In this sense, the FBST method enables business researchers to provide converging evidence that results obtained in laboratory experiments are replicable in the real world and actionable by managers.

3.3. Data quality control

In that the efficacy of an experiment is only as good as the quality of the data it produces, the goal of an experimental research design is to enhance the conditions that promote natural responses and depreciate the occurrence of artificial or unnatural participation ([Zikmund, Babin, Carr, & Griffin, 2013](#)). FBST experiments return high quality data for four reasons. First, they occur in a natural setting (i.e., during users’ normal usage of Facebook as a social media) where participants are unaware of the hypotheses being tested. This feature reduces the risk of experimenter effects compared to laboratory experiments, making FBST comparable to MTurk experiments (cf. [Paolacci et al., 2010](#)). Second, the naturalistic conditions and lack of incentives for participants also eliminate the risk of demand biases, speeding and cheating, which are traditionally high in MTurk experiments ([Ford, 2017](#)), but can also occur in laboratory settings. That is, while participants of laboratory and MTurk experiments are aware they are involved in a research study, FBST experiments occur in natural settings, with users viewing ads as part of their normal Facebook usage, unaware a split testing

experiment is being conducted. Third, Facebook’s single-user login ensures that an experimental stimulus is shown only once to the same user. This quality control feature limits the risk of multiple response-clicks as long as researchers ensure, through the dedicated filter, that the number of *impressions* or times an advertisement is shown equals the *reach*, or the total number of users that see the advertisement at least one time. We note how a single-user does not necessarily equate to a single-participant, as in rare cases individuals may possess multiple accounts. However, apart from business accounts, Facebook policy permits only one personal account precisely because they need to guarantee their clients (i.e., advertisers) legitimacy for the number of users reached by the ads ([Facebook, 2019](#)). Fourth, single-user login and strict account policies also limit the presence of fake accounts, avoiding the increasingly common issue of automated bots filling in surveys on MTurk ([New Scientist, 2018](#); [Wired, 2018](#)).

In summary, FBST experiments yield higher data quality compared to laboratory and MTurk experiments, because the risk of experimenter effects is lower than laboratory settings and comparable to MTurk, the risk of multiple responses by the same user is lower than MTurk and comparable to laboratory settings, the risk of speeding and cheating is lower than both laboratory and MTurk settings, and the risk of automated bot contamination is lower than MTurk.

3.4. Sample representativeness

Sample representativeness is a critically important condition of experimental research in that researchers want to avoid over- or under-representing certain opinions, thereby creating a sample bias that could undermine the external validity of the data ([Zikmund et al., 2013](#)). Compared to alternative experimental settings, FBST experiments have higher external validity due to their ability to reach large subsamples of the general population ([Matz et al., 2017](#)). Facebook requires its users to provide personal information in terms of age and gender, and tracks location, allowing researchers to target audience segments according to needs. However, we note one key limitation in that Facebook does not verify information provided by users (e.g. age, gender, education levels and political orientation). Only page engagements, online behaviors, and interests recorded through user Likes are recorded by the platform. In this sense, the validity of user information is superior to MTurk, in which all information is self-reported and prone to falsification. It was also comparable to laboratory settings, in which information such as age, gender, and education levels could be verified, but other, including interests, hobbies, and dispositions are entirely self-reported. Access to a large population and a sophisticated targeting capability also reduce susceptibility to coverage error and sample heterogeneity across laboratories (see [Paolacci et al., 2010](#)).

This is not to say that we believe the sample representativeness of FBST experiments is necessarily high. Previous studies in the US and UK report that social media users tend to be younger and better educated than non-users ([Greenwood et al., 2016](#); [Mellon & Prosser, 2017](#)). Studies conducted in Canada, on the other hand, found that the 35–74 years old Facebook segment found Facebook was representative of the general Canadian population ([Shaver et al., 2019](#)). More research on the *absolute* sample representativeness of Facebook users is direly needed, especially across different countries. What we can affirm, however, is that experiments conducted through the FBST have a higher sample representativeness *relative* to laboratory and MTurk recruiting methods, and a comparable representativeness relative to specialized providers of online panels ([Smith, Roster, Golden, & Abaum, 2016](#)).

4. Setting up a FBST experiment

To support the experimental efforts of business researchers, we now provide a step-by-step guide on how to set up a simple experimental design through the (i) campaign (i.e., objectives and selection of A/B split testing), and (ii) ad set (i.e., variable tested, targeting procedure,

budget and scheduling) tabs.

4.1. Campaign

The first step for a FBST experiment is to choose the campaign objective. While advertisers may be driven by specific objectives such as increasing conversion rates, store traffic, or app installs, researchers simply aim to randomly assign experimental conditions to an even split of individuals who only see an experimental stimulus once. To do so, in the FBST interface:

1. Select “Reach” as the marketing objective and provide a name for the experiment. This option will ensure the stimuli are shown to an even split of the highest possible population that your budget allows. While other options are available, they either introduce ad repetition effects (“brand awareness”) or employ optimization algorithms aimed at increasing conversion rates;
2. Below the campaign name, select “Create A/B test”;
3. Do not select “Campaign Budget Optimization” to avoid confounds from delivery optimization algorithms.

4.2. Ad set

In the ad set pane, researchers need to set up a landing page where users can land after clicking on the ad, select the independent variable they want to test, provide information on the audience they wish to reach, ensure delivery optimization options are not selected, and indicate the research budget they wish to allocate. We indicate in italics the names of each tab as used in the FBST interface.

4. *Page*. Creating a Facebook page provides users with a place to land after clicking on the experimental stimulus. While this tutorial uses click-through as the dependent variable, other designs may wish to capture page engagements, page likes, and even online purchases or subscriptions. All of these dependent variables require the development of a landing page whose content is aligned to the experimental stimulus;
5. *Variable*. Select “Creative” as the variable to test. This will enable A/B testing in the following pane;
6. *Audience*. Select which audience will be part of the experiment. Location, age, gender, and language are the default options. Only location needs to be specified, and if no modifications are implemented, the campaign will run across all genders in the age range 18–65+ (note that Facebook allows for the selection of younger audiences, as young as 13 years). More sophisticated targeting options include other demographics (e.g., education, financial status, relationship status), interests (e.g., entertainment, sports and outdoors), and behaviors (e.g., digital activities, purchase behaviors), each undergirded by more precise subsections. These targeting variables can lead to even more precise tests when a specific audience is the focus of the research. Audience profiles can also be saved through the corresponding functionality for use in future experiments. The caveat here is that the more stringent the audience requirements, the lower the maximum potential reach of the experiment. For instance, at the time this tutorial was written, selecting a U.S. female audience aged 18–40 returned a maximum potential reach of 61 million users. Adding the additional targeting variable “College graduate” reduced this number to 18 million. In our demonstration, our targeting procedure was to select a U.S.-only audience with an age range of 18–65+, which led to a total reach of 142,262 users.
7. *Placement*. Leave the “Automated Placement” option.
8. *Delivery optimization*. First, ensure that “Optimization for Ad Delivery” indicates “Reach”. This will ensure consistency with the experiment objective of simply randomly assigning as many individuals as the budget allows. Second, ensure that “Frequency Cap”

indicates “1 impression every X days”, where $X = 1 + \text{number of days the experiment is running}$. This will ensure each individual is assigned to one condition only once, avoiding repetition effects.

9. *Split test budget & schedule*. The total reach of the experimental campaign is a function of the total budget evenly split across experimental conditions. FBST asks the researcher to indicate a daily budget that will be split across conditions. The typical time frame employed by prior research is seven days (Matz et al., 2017), with Facebook recommending a minimum of four days to enable meaningful tests. When it comes to budgeting decisions, FBST will provide an estimate of the daily reach the allocated budget affords. Prior research reports click-through rates ranging from 0.001% to 0.013% (Matz et al., 2017), so we recommend a budget allocation that will guarantee a total reach of 100,000 users to ensure an aggregated minimum of 100 click-throughs pooled across the two conditions.

5. Preparing the experimental stimuli and running the experiment

Once the steps detailed for the “campaign” and “ad set” panes have been completed, researchers are required to upload the experimental stimuli they wish to test. Before this can happen, however, it is necessary to first develop stimuli that clearly operationalize the constructs of interest, and then to pre-test them to ensure the validity of the constructs manipulated and embedded in the stimuli. Because FBST runs on an ecologically valid delivery platform (i.e., Facebook), experimental manipulations need to fit the format of realistic visual, textual, or audio-visual advertisements.

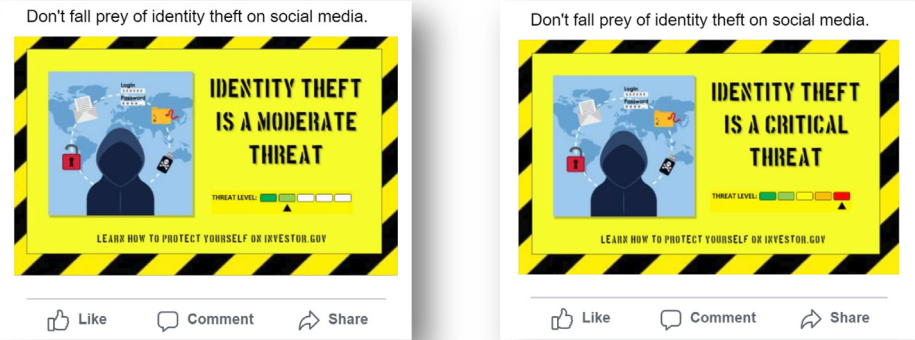
5.1. Theory-driven manipulations: an illustration based on protection motivation theory

For the sake of illustrating the complete process of running a FBST experiment, we aim to replicate the results of an experimental design testing the effectiveness of different fear appeals on users’ decision to adopt desirable information security behaviors (Johnston & Warkentin, 2010; Orazi & Pizzetti, 2015). Fear appeals are commonly used to motivate adherence with behaviors that can prevent or nullify impending threats (Johnston & Warkentin, 2010; Orazi, Warkentin, & Johnston, 2019). Fear appeals are used in a variety of business domains, including insurances (LaTour & Zahra, 1988), pharmaceuticals (Wakefield et al., 2005), and cybersecurity (Johnston et al., 2015), as well as several public service announcements and public policy campaigns that promote protective behaviors (e.g., sunscreen use: Passyn & Sujan, 2006) and invite audiences to refrain from compulsive behaviors (e.g., problem gambling: Orazi, Bove, & Lei, 2015).

Protection Motivation Theory (PMT: Rogers, 1975; Maddux & Rogers, 1983) is the dominant theoretical framework when it comes to understanding the persuasiveness of fear appeals. While a review of PMT is beyond the scope of this methodological paper (for reviews, see Witte & Allen, 2000; Boss et al., 2015), PMT’s core tenet implies that messages perceived to be more threatening are more persuasive than messages perceived to be less threatening and the severity of the depicted threat is one of the key components that influence how users assess how threatening the message is. Accordingly, the likelihood a user will adopt a protective behavior increases when the user perceives the depicted threat to be severe. This effect of threat severity on protection motivation has been confirmed in several meta-analyses (de Hoog et al., 2007; Witte & Allen, 2000) and consistently replicated in marketing (Orazi & Pizzetti, 2015) and information security research (Johnston et al., 2019). The focal hypothesis we aim to test in our experimental demonstration is as follows:

H1: High-threat messages will increase protection motivation more so than moderate-threat messages.

Table 2
Single factor design: Empirical demonstration stimuli and results.

	<i>Moderate threat ad</i>	<i>High threat ad</i>
		
Reach	69,645	72,619
Gender	Female: 47%, Male: 53%	Female: 45%, Male: 55%
Age	18-24 = 40,852 25-34 = 18,179 35-44 = 3,418 45-54 = 1,664 55-64 = 1,920 65+ = 2,560	18-24 = 41,732 25-34 = 19,133 35-44 = 3,360 45-54 = 1,774 55-64 = 2,016 65+ = 3,321
Impressions	69,874	72,620
Frequency	1.0033	1.0000
Link clicks	87 [86]	202
Expenditure	\$163.84	\$163.76
CTR	0.128	0.278
CPC	\$1.88	\$0.81

Note: Click-through = (clicks/reach)*100; CTR = click-through rate; CPC = expenditure/clicks; [clicks in brackets indicate correction by frequency]

Note that we compare high vs. moderate threat messages (instead of high vs. low) to increase verisimilitude in the field setting (i.e., why would the government promote protection motivation if a threat is not severe in the first place?) and provide a stronger test than previous research comparing high vs. low threat messages. In our empirical demonstration of the FBST, we developed the stimuli following the guidance of previous research on information security (Johnston & Warkentin, 2010; Johnston et al., 2015). Specifically, we designed two simple fear appeals which only differ in their level of threat severity (moderate vs. high) and invited users to click through to access a government web page providing cybersecurity recommendations (see Table 2 for the stimuli employed).

5.2. Pre-testing manipulations and embedding them in the ad A/B pane

After the experimental stimuli have been developed, it is necessary to rigorously pre-test the validity of the experimental manipulation (see Perdue & Summers, 1986, for more details on independent sample pre-tests). While the FBST does not allow for implementing manipulation checks, the availability of affordable crowdsourcing platforms can help overcoming this issue. For instance, researchers interested in comparing the effects of moderate versus high threat severity fear appeals on protection motivation may wish to use MTurk or an online panel data provider to pre-test whether the high threat severity message leads to significantly higher perceptions of threat severity in comparison to the moderate threat message. In case the manipulation checks are not successful, a panel of users can also provide qualitative feedback that may help hone the fear appeal's message to its appropriate level of threat severity.

Continuing the illustration through our fear appeal example, after applying a filtering procedure to minimize the impact of automated bots and speeders (% HIT acceptance: 98; # HITS: 500), we recruited one-hundred-and-five MTurk workers ($M_{\text{age}} = 36.52$, $SD = 10.72$, 45.7% female) for a one-minute pictorial evaluation task. These participants were compensated \$0.30 and randomly assigned to a simple (threat: moderate vs. high) between-subject design before completing measures of perceived threat severity on a 3-item ("The threat depicted is: serious/severe/significant"), 7-point Likert scale (Johnston & Warkentin, 2010). After averaging the items in a single indicator of threat severity ($\alpha = 0.95$), we ran an independent sample *t*-test to ensure the validity of our manipulation.

Results confirmed that the high threat condition was perceived as significantly more severe ($M = 6.13$, $SD = 0.85$) than the moderate threat condition ($M = 4.46$, $SD = 1.42$ $t(103) = 7.23$, $p < .001$). Once the pre-test manipulation checks confirm the validity of the experimental manipulations, researchers can include them in the FBST as the focal stimuli that will be randomly assigned to the target audience. Procedural steps 10–13, necessary to embed the experimental stimuli in the FBST platform, are detailed in Web Appendix A.

6. Data analysis

6.1. Data export: aggregated output, key metrics, and their meaning

The campaign will run for as many days as indicated at point 9. Once the campaign is over, the FBST will provide exportable output in the "Ad Reporting" tab of Facebook Ad Manager. While the report can be personalized using several advertising-specific metrics, we

recommend selecting the following performance metrics to both provide an ecologically valid assessment of the campaign results, as well as the key count metrics necessary to perform non-parametric tests of significance.

Reach: the number of unique users that were exposed to each experimental condition. Total reach should be the same across all experimental conditions.

Impressions: the number of times an experimental condition was shown. Impressions should equal the reach to avoid repetition effect.

Frequency: the ratio between impressions and reach. Frequency should ideally be equal to one. In our tests, occasional algorithm errors returned slightly higher values even when indicating a maximum of one single exposure per participants (e.g. Frequency = 1.01).

Amount spent: the total cost for the campaign, split across conditions. This amount should be identical for all experimental conditions.

Clicks: the number of clicks for each experimental condition which represent a proxy of users' engagement and motivation. This is the key metric employed in non-parametric tests of significance.

Click-through rate (CTR): the ratio between total clicks and total impressions, which provides a quick indicator of the effectiveness of one condition over another in motivating to click-through to the landing page.

Cost-per-click (CPC): the ratio between the amount spent in one experimental condition and the total clicks generated by that condition. Because the amount spent is constant across conditions, comparatively lower amounts of CPC in one condition can be used as a proxy for effectiveness. However, CPC is only relevant for capturing higher cost efficiency for one condition over the alternative(s) and should not be employed in non-parametric tests for significance. In our empirical demonstration, we use CTR as the focal dependent variable.

6.2. Data analysis: non-parametric tests and logistic regression on simulated data

While the values of both CTR and CPC immediately provide managerial insights as to whether one condition is more effective than the other(s), researchers still need to rely on statistical tests to establish whether the difference is statistically significant. The data provided in the output file, however, is at an aggregate level, which prevents inferential statistics tests. To overcome this limit, we suggest two approaches: (i) a non-parametric test, specifically a chi-square comparison using click counts, or (ii) a logistic regression analysis on a simulated dataset based on click-counts, coding 0 for non-clicks and 1 for clicks for each condition. We illustrate both approaches below.

Chi-square comparison on clicks. If the reach is similar across conditions, then researchers can divide based on clicks and non-clicks and perform a chi-square test. In our empirical demonstration, we obtained a total reach of 142,262. Using a chi-square comparison, we compared the proportion of users who clicked versus did not click the ad in both experimental conditions. As displayed in Table 2, the moderate threat condition attracted 87 clicks (and correspondingly 69,558 non-clicks) and the high threat condition attracted 202 clicks (and correspondingly 72,417 non-clicks).

Clicks and non-clicks can be included as columns, and low versus high threat as rows in a cross-tab, to calculate the expected and observed frequencies for a chi-square test aimed at assessing whether the distribution of the outcomes depends on the experimental manipulation. A review of how to perform a chi-square test is beyond the scope of this tutorial, but we note that several online calculators allow researchers to perform this test automatically (we recommend the intuitive interface of www.socscistatistics.com/tests/chisquare2/default2.aspx, Social Science Statistics, 2019). Imputing the values provided in the example above, a chi-square test with one degree of freedom (i.e., $[rows-1] * [columns-1]$) confirmed a significant difference in the proportion, $\chi^2(1) = 682,886.81$, $p < .001$, such that the high threat severity ad resulted in higher click-through (0.278)

compared to the moderate threat condition (0.128). This result means that the high threat severity condition was more effective in attracting users' clicks, which were used as a proxy for protection motivation.

The CPC can also be used as a managerially relevant indicator of cost efficiency. The CPC was \$0.81 for the high threat severity ad and \$1.88 for the moderate threat severity ad, indicating higher effectiveness, and thus cost-efficiency, for the high threat severity ad. Table 2 summarizes the results and key metric employed. On a concluding note, chi-square tests are sensitive to sample size and the large amount of data produced by FBST tend to inflate the chi-square value. Bergh (2015) provides procedural recommendations on how to adjust chi-square values with large sample sizes.

Logistic regression on simulated data. A second approach to data analysis involves the creation of a dataset simulating the clicks and non-clicks across conditions. In a simple A/B split design comparing two conditions, this entails creating (i) a column coding for the manipulation [0 = A; 1 = B], and (ii) a column coding for clicks [0 = No; 1 = Yes]. In our example, this will mean that the first column codes 0 69,645 times and 1 72,619 times. Next, the second column codes 0 69,558 times and 1 87 times for all rows that display a 0 in the first column (condition A), and 0 72,417 times and 1 202 times for all rows that display a 1 in the first column (condition B). This simulated dataset enables running a logistic regression that will return a significance test for the odds ratios, or how more likely to click are participants assigned to condition B, compared to those assigned to condition A. In our example, the logistic regression similarly returns a positive effect for the high threat condition on number of clicks, such that users exposed to high threat severity are 2.23 times more likely to click through the ad than those exposed to the low threat severity condition, ($B = 0.802$, $SE = 0.13$, $Wald = 39.05$, $p < .001$, $Exp(B) = 2.23$). We advise the use of this analytical method when sample sizes are larger than 10,000 observations per cell and chi-square tests return inflated values. As a robustness check, we conducted a replication of the FBST experiment herein reported using MTurk as the data source. Results are available in Web Appendix C.

Correcting for algorithmic discrepancies. In some instances, the FBST will show the same ad to a user twice even when experimenters indicate a maximum of one impression per user reached in the "Frequency Cap" discussed under the *Delivery optimization* phase. This algorithmic error is endogenous to the platform and needs to be accounted for. As a correction rule, we suggest to divide the total number of clicks by the frequency, rounding down. In our empirical demonstration, this correction results in the removal of one click from the moderate threat condition (i.e., $87/1.0033 = 86.71$). This correction does not alter the significance and interpretation of the results.

7. Designing 2 × 2 experiments using the FBST

The experimental design described and demonstrated in the previous section can be easily implemented for single factorial and 2 × 2 designs where both independent variables are manipulated. For 2 × 2 designs, for instance, researchers can create four ads and then test interaction effects using the logistic regression approach on simulated data described in 6.2 *Data analysis: Non-parametric tests and logistic regression on simulated data*.

At times, however, researchers may be interested in testing 2 × 2 designs in which one factor is manipulated and the other is measured. Facebook maintains an extensive database of deidentified users' information clustered in three groups: *demographics* (e.g., education level, income), *interests* (e.g., entertainment, fitness), and *behaviors* (e.g., operating systems used, mobile accesses). These groups can be accessed under the *Detailed Targeting* pane when selecting the audience (see 4.2 *Ad set*, point 6). Most demographic and behavioral targeting variables can be easily implemented in a 2 × 2 design as measured factors because they are non-overlapping. For demographics, a user will list the highest level of education achieved and will fall within a specific

income bracket. For behaviors, a user interacting with the platform at any given time will do so through one operating system and one device. An important caveat pertains interests, which can overlap. For example, a user can be interested in both entertainment and fitness and if these were the targeting variables used, then the same user could be exposed to two campaigns and confound the results (see *Limitations and future research* for related future advancements). This issue is resolved by constraining measured factors to demographics and behaviors. Because most demographic and behavioral variables are non-overlapping (i.e., a user cannot list both “some high school” and “doctoral degree”: only the highest level of education is listed), they allow to run parallel campaigns without sample contamination. Below, we provide an empirical illustration of this method.

7.1. Theoretical foundations: an illustration based on terror management theory

We illustrate how to set up and run a 2×2 design with one measured factor using education as the demographic variable and Terror Management Theory or TMT (Greenberg, Pyszczynski, & Solomon, 1986) as the theoretical framework. Consider how highly educated individuals are more sceptical and less likely to click on advertisements. How to soften this advertising barrier? According to TMT, awareness of one own's mortality triggers feelings of existential anxiety. To defend against this aversive feeling, individuals use personal standards and enduring conceptions, broadly captured as cultural worldviews, as means by which to feel they are special creatures endowed with purpose, meaning, and significance, rather than “mere animals fated to absolute annihilation when they die” (Maxfield et al., 2007: 342). Accordingly, reminders of mortality make cultural worldviews more salient. Because cultural worldviews are structured and enduring conceptions of reality (Maxfield et al., 2007), more educated individuals should possess more articulated worldviews. Subtle reminders of mortality should thus have greater effectiveness for more (vs. less) educated individuals as they activate their cultural worldviews. When the mortal reminder is accompanied by a solution (e.g., a PN2 mask against an infectious disease), we expect the heightened activation of cultural worldviews to translate in greater acceptance of the solution, in our context clicking on the ad proposing the solution. In summary:

H2: For more (vs. less) educated individuals, high (vs. low) mortality salience will increase click-through behaviors.

7.2. 2×2 experimental designs

To set up a 2×2 design in which one factor is manipulated and the other is measured, researchers can follow the same procedure illustrated for single-factor designs with one exception: researchers need to create two identical FBST campaigns with different audiences, one for each of the non-overlapping measured factors (i.e., in our example, education). The distinctive component of these otherwise identical campaigns is the measured factor. To illustrate, we were interested in understanding how the level of education amplifies perceptions of mortality salience, leading to increased click-through behavior. For the first campaign (i.e., low education level), we selected Australian adults of age 25–65+ that reported “some high school” as their level of education. For the second campaign (i.e., high education level), we selected Australian adults of age 25–65+ that reported “doctoral degree”, “graduate school” or “professional degree (MD)” as their level of education. We chose these education levels because they represented the lower and upper limits of the variable education in Facebook. We also included only users above the age of 25 to exclude young individuals currently enrolled in high school and only include adults whose highest education level was “some high school”.

We developed the stimuli following similar steps as in our first empirical demonstration. As the context of investigation, we chose the

SARS-CoV-2 outbreak of January 2020² to ensure users will be responsive to the campaigns. Our stimuli portrayed a PN2 mask with two different copy texts. The low mortality salience condition stated “Protect yourself from the coronavirus with a PN2 mask”, whereas the high mortality salience condition stated “Fear the contagion? Protect yourself from the deadly coronavirus with a PN2 mask” (see Table 3 for the stimuli). We pre-tested the two stimuli to ensure they effectively manipulated mortality salience (see Web Appendix D for the pre-test results). The two campaigns ran parallelly for four days from the 4th to the 7th of February 2020, with a total budget of \$240 (\$15* 4 conditions*4 days).

7.3. Data analysis and discussion

A chi-square comparison was initially used to compare the proportion of users who clicked an ad versus users who did not across all four conditions. Results confirmed a significant difference in the proportion ($\chi^2(3) = 9.62, p < .002$). To decompose simple effects, we first compared the proportion of clicks versus non-clicks in the low mortality salience (Non-clicks = 17,920 vs. Clicks = 94) and high mortality salience conditions (Non-clicks = 17,913 vs. Clicks = 137) for high education users, finding a significant difference ($\chi^2(1) = 7.97, p < .005$). We then compared the proportion of clicks versus non-clicks in the planning (Non-clicks = 17,982 vs. Clicks = 258) and coincidence conditions (Non-clicks = 18,266 vs. Clicks = 228) for low education users, finding no significant difference ($\chi^2(1) = 2.32, p = .128$). As hypothesized, users with higher education levels, while generally more resistant to click online ads, were more likely to click ads that primed high mortality salience in comparison to ads that primed low mortality salience. Low education users did not significantly differ in their reactions to the two types of ad. Table 3 summarizes the results.

8. Discussion

This primer provides a cost-effective recruiting method to conduct online field experiments with high external validity. We have discussed the implementation of the FBST for single-factor designs and 2×2 designs with both factors manipulated, or one factor manipulated and the other measured. As FBST experiments possess high ecological validity at the independent variable, delivery platform, and dependent variable levels, they are complementary to other experimental design settings. By no means have we intended to indicate FBST as a substitute to other experimental settings such as those provided by laboratory studies and MTurk, which possess high internal validity. When the objective is to provide externally valid evidence, FBST represents a robust and affordable way to conduct field experiments in an online setting, and one that will greatly assist researcher able to embed their manipulated variables into textual, visual, and audio-visual experimental stimuli. In this concluding section, we discuss the applicability of the FBST beyond advertising testing, some theoretical implications of our findings, and future methodological refinements.

8.1. Applicability beyond advertising testing

Online field experiments using Facebook as the delivery platform have historically been characterized as having reasonably high levels of external validity but lacking internal validity due to a number of challenges related to a lack of random assignment (Eckles et al., 2018) and biases associated with delivery optimization algorithms (Gordon et al., 2018). Given the pressure business researchers are facing to

² At the time the study was conducted, the SARS-CoV-2 was not yet declared a pandemic by the WHO and was affecting mainly China. Conclusive evidence on how the virus propagated was not yet available.

Table 3

2 × 2 designs: Empirical demonstration stimuli and results.

No mortality salience

Mortality salience

Protect yourself from the coronavirus with a PN2 mask.

Like

Comment

Share

Fear the contagion? Protect yourself from the deadly coronavirus with a PN2 mask.

Like

Comment

Share

Low Education

High Education

Low Education

High Education

Reach	18,240	17,788	18,188	18,050
<i>Gender</i>	F: 45%, M: 55%	F: 44%, M: 56%	F: 43%, M: 57%	F: 44%, M: 63%
<i>Age</i>	25-34 = 6,438	25-34 = 6,360	25-34 = 6,448	25-34 = 6,684
	35-44 = 4,125	35-44 = 3,972	35-44 = 4,092	35-44 = 4,137
	45-54 = 3,676	45-54 = 3,604	45-54 = 3,664	45-54 = 3,550
	55-64 = 2,321	55-64 = 2,255	55-64 = 2,373	55-64 = 2,261
	65+ = 1,417	65+ = 1,597	65+ = 1,571	65+ = 1,418
Impressions	18,240	18,014	18,494	18,050
Frequency	1.0000	1.0127	1.0168	1.0000
Link clicks	258	94 [93]	228 [224]	137
Expenditure	\$57.43	\$57.45	\$57.46	\$57.42
CTR	1.414	0.521	1.232	0.759
CPC	\$0.22	\$0.61	\$0.25	\$0.42

Note: Click-through = (clicks/reach)*100; CTR = click-through rate; CPC = expenditure/clicks; [clicks in brackets indicate correction by frequency].

conduct research that possesses both high degrees of external *and* internal validity, FBST is a welcomed addition to the Facebook delivery platform that not only negates these challenges, but also provides a number of advantages that make it attractive for conducting research outside of traditional advertisement testing.

As an example, consider the study of highly nuanced phenomena such as the study of online behaviors resulting from deepfake simulations. A deepfake occurs when human images are superimposed with artificial intelligent driven manipulations of the images to create nearly indistinguishable human representations. Because the success of deepfakes is dependent on the extent to which they mimic real human personas, the study of them requires a testing environment, treatments, and outcomes that have a high degree of ecological validity. If any one of these experimental design elements lack ecological validity, the entire study will suffer accordingly and its results will be questionable. Because FBST is flexible and able to embed theoretical constructs operationalized as text, visual, or audiovisual elements as manipulated independent variables in an experiment, business researchers can purpose it toward the study of online behaviors where familiarity with and previous exposure to a particular simulation, message, or visual is relatively limited and highly contextual (e.g., deepfake simulations).

Beyond its usefulness as a platform for the study of highly nuanced phenomena, FBST is also applicable to the study of rhetorical discourse and its influence on various audience segments. Rhetorical discourse is used by organizational managers to persuade individuals or audiences to adapt to the ideals and values of the organization and its leaders (Hartelius & Browning, 2008). Because of its ability to randomly assign participants to experimental conditions, FBST makes it possible to study

how various forms of rhetorical discourse influence audience behaviors. In the information security context, the understanding of how best to design the rhetorical discourse contained in a data breach disclosure statement is critical to a chief information officer's (CIO) or chief information security officer's (CISO) ability to maintain the confidence of his or her customers and constituents. It is also critically important to the success of a politician seeking office, where certain online behaviors by constituents in social platforms such as Facebook are often seen as proxies for support or non-support.

8.2. Implications of empirical studies

While the principal aim of this paper was to provide a primer to conduct FBST experiments, the two field experiments we conduct to demonstrate how to implement FBST designs provide implications to both protection motivation theory (Maddux & Rogers, 1983; Rogers, 1975) and terror management theory (Greenberg, Pyszczynski, & Solomon, 1986). The implications of the FBST method for information security research can most likely be seen in its ability to implicitly force business researchers to consider ecologically valid fear appeals in their studies. Fear appeals are effectively treatments intended to serve as catalysts for behavior change, and recent research involving fear appeals in the information security literature has suggested that security scholars have undervalued the role of fear appeals as drivers of security behaviors (Orazi et al., 2019). In particular, little attention has been paid to the importance of the rhetorical validity of fear appeals; rhetorical validity referring to a specialized form of ecological validity in which the language used in the appeal is consistent with the threat

environment and expectations of its audience (McKerrow, 1977).

This undervaluing of fear appeals and their lack of rhetorical validity has contributed to a stagnation in the advancement of fear appeals and fear appeal theories within the information security research community (Johnston et al., 2015). The advantages of FBST described and demonstrated in this primer could help ensure that fear appeals are prepared and tested in a manner that ensures they are rhetorically valid and more efficacious in motivating security behaviors among their audiences. By leveraging FBST in fear appeal experiments, information security scholars can break the habit of undervaluing the importance of fear appeal treatments' rhetorical validity and start to reverse the trend of questionable fear appeal-induced outcomes. Our empirical demonstration offers a first step in this direction, providing evidence for the robustness of the FBST methodology. We find support for a superior effect of high-threat (vs. moderate-threat) fear appeals on protection motivation, conceptually replicating the results of prior marketing (Orazi & Pizzetti, 2015) and information security studies (Johnston & Warkentin, 2010). To ensure the robustness of our results, we also directly replicate our empirical demonstration through an online experiment using Amazon Mechanical Turk, again finding the same results.

Our empirical demonstration of a 2×2 design based on terror management theory also yields interesting theoretical insights. We demonstrate how individuals with more complex and articulated worldviews reality (Maxfield et al., 2007), for instance because of higher education, are more susceptible to high (vs. low) mortality salience. High mortality salience makes cultural worldviews more salient as well in an attempt to defend against the aversive feeling of existential anxiety. Our results indeed demonstrate in an ecologically valid setting that users with higher education were indeed more likely to click an advertisement of a protective mask when the ad primed high mortality salience.

8.3. Future methodological developments

Despite the merits we advocate, the FBST would benefit greatly from research work that aims to extend the use of measured variables as proxies for psychological constructs. As discussed, users' demographics, interests, and behaviors are recorded by Facebook as deidentified information, but can be accessed as targeting variables to implement 2×2 design.

Understanding which variables available in the FBST targeting variable pane can be used as proxies of psychological constructs represents a very fruitful area for future research. Early Facebook studies that predate the FBST methodology were mainly focused on psychological persuasion—a targeting approach aimed at maximizing message effectiveness by tailoring persuasive communications to individual traits and dispositions (Matz et al., 2017). To this end, *myPersonality.org*, developed and maintained by David Stillwell and Michal Kosinski, allowed researchers to access a comprehensive dataset of anonymized data linking personality traits to Facebook Likes, which became effective predictors of personality. As stated on the website, the creators stopped sharing the data in April 2018 due to administrative burden and increased regulatory pressure.

We exhort further research aimed at establishing correlations between Facebook targeting variables and psychological construct relevant to business research. One avenue to explore is whether socio-demographic variables and preferences can be used as proxies for dispositional variables of interest based on existing literature. For instance, a user's occupation can be indexical of need for cognition (Cacioppo & Petty, 1982). A research paper investigating whether occupation can be used robustly as a proxy for need for cognition would ideally (a) establish additional correlational evidence between the two variables, (b) provide evidence for the focal effect by measuring need for cognition using laboratory or web experiments, and (c) replicate the results using occupation as the targeting proxy through a FBST experiment. Research

efforts in this direction would provide a substantive contribution to the advancement of the method.

Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jbusres.2020.06.053>.

References

- Bell, E., Bryman, A., & Harley, B. (2018). *Business research methods*. Oxford University Press.
- Bergh, D. (2015). Chi-squared test of fit and sample size-A comparison between a random sample approach and a chi-square value adjustment method. *Journal of Applied Measurement*, 16(2), 204–217.
- Boss, S., Galletta, D., Lowry, P. B., Moody, G. D., & Polak, P. (2015). What do systems users have to fear? Using fear appeals to engender threats and fear that motivate protective security behaviors. *MIS Quarterly*, 39(4), 837–864.
- Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal*, 5(4), 437–474.
- Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, 42(1), 116–131.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297.
- Daly, T. M., & Natarajan, R. (2015). Swapping bricks for clicks: crowdsourcing longitudinal data on Amazon Turk. *Journal of Business Research*, 68(12), 2603–2609.
- De Hoog, N., Stroebe, W., & De Wit, J. B. (2007). The impact of vulnerability to and severity of a health risk on processing and acceptance of fear-arousing communications: A meta-analysis. *Review of General Psychology*, 11(3), 258–285.
- Eckles, D., Gordon, B. R., & Johnson, G. A. (2018). Field studies of psychologically targeted ads face threats to internal validity. *Proceedings of the National Academy of Sciences*, 115(23), E5254–E5255.
- Facebook (2019) Split testing. Available at <https://www.facebook.com/business/help/1738164643098669>. Accessed March 27, 2019.
- Ford, J. B. (2017). Amazon's Mechanical Turk: a comment. *Journal of Advertising*, 46(1), 156–158.
- Gordon, B. R., Zettermeyer, F., Bhargava, N., & Chapsky, D. (2018). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook, working paper.
- Greenberg, Jeff, Pyszczynski, Tom, & Solomon, Sheldon (1986). *The causes and consequences of a need for self-esteem: A terror management theory*. Public self and private self. New York: Springer 189–212.
- Greenwood, S., Perrin, A., & Duggan, M. (2016). Social Media Update 2016. Pew Research Center. Available at: http://assets.pewresearch.org/wp-content/uploads/sites/14/2016/11/10132827/PI_2016.11.11_Social-Media-Update_FINAL.pdf.
- Hartelius, E. J., & Browning, L. D. (2008). The application of rhetorical theory in managerial research: A literature review. *Management Communication Quarterly*, 22(1), 13–39.
- Inman, J. J., Campbell, M. C., Kirmani, A., & Price, L. L. (2018). Our vision for the Journal of Consumer Research: It's all about the consumer. *Journal of Consumer Research*, 44(5), 955–959.
- Johnston, A. C., & Warkentin, M. (2010). Fear appeals and information security behaviors: An empirical study. *MIS Quarterly*, 34(3), 549–566.
- Johnston, A. C., Warkentin, M., & Siponen, M. (2015). An enhanced fear appeal framework: Leveraging threats to the human asset through sanctioning rhetoric. *MIS Quarterly*, 39(1), 113–134.
- Johnston, A. C., Warkentin, M., Siponen, M., & Dennis, A. (2019). Speak their language: A strategy for persuasive message fit in support of information security behavioral compliance. *Decision Sciences*, 50(1), 245–284.
- LaTour, M. S., & Zahra, S. A. (1988). Fear appeals as advertising strategy: Should they be used? *Journal of Services Marketing*, 2(4), 5–14.
- Lewis, R. A., Rao, J. M., & Reiley, D. H. (2011, March). Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th international conference on World wide web* (pp. 157–166). ACM.
- Maddux, J. E., & Rogers, R. W. (1983). Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *Journal of Experimental Social Psychology*, 19(5), 469–479.
- Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48), 12714–12719.
- Maxfield, M., Pyszczynski, T., Kluck, B., Cox, C. R., Greenberg, J., Solomon, S., & Weise, D. (2007). Age-related differences in responses to thoughts of one's own death: mortality salience and judgments of moral transgressions. *Psychology and Aging*, 22(2), 341–353.
- McKerrow, R. E. (1977). Rhetorical validity: An analysis of three perspectives on the justification of rhetorical argument. *The Journal of the American Forensic Association*, 13(3), 133–141.
- Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3), 1–9.
- New Scientist (2018). Bots on Amazon's Mechanical Turk are ruining psychology studies. Available at: <https://www.newscientist.com/article/2176436-bots-on-amazons->

- mechanical-turk-are-ruining-psychology-studies/. Accessed March 27, 2019.
- Orazi, D. C., Lei, J., & Bove, L. L. (2015). The nature and framing of gambling consequences in advertising. *Journal of Business Research*, 68(10), 2049–2056.
- Orazi, D. C., & Pizzetti, M. (2015). Revisiting fear appeals: A structural re-inquiry of the protection motivation model. *International Journal of Research in Marketing*, 32(2), 223–225.
- Orazi, D. C., Warkentin, M., & Johnston, A. C. (2019). Integrating construal-level theory in designing fear appeals in IS security research. *Communications of the Association for Information Systems*, 45(1), 397–410.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5), 411–419.
- Passyn, K., & Sujan, M. (2006). Self-accountability emotions and fear appeals: Motivating behavior. *Journal of Consumer Research*, 32(4), 583–589.
- Perdue, B. C., & Summers, J. O. (1986). Checking the success of manipulations in marketing experiments. *Journal of Marketing Research*, 23(4), 317–326.
- Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change. *The Journal of Psychology*, 91(1), 93–114.
- Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, 12(2), 225–237.
- Shaver, L. G., Khawer, A., Yi, Y., Aubrey-Bassler, K., Etchegary, H., Roebbothan, B., & Wang, P. P. (2019). Using facebook advertising to recruit representative samples: Feasibility assessment of a cross-sectional survey. *Journal of Medical Internet research*, 21(8), 1–15.
- Smith, S. M., Roster, C. A., Golden, L. L., & Albaum, G. S. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, 69(8), 3139–3148.
- Social Science Statistics (2019). Chi-Square Test Calculator, Available at: www.socscistatistics.com/tests/chisquare2/default2.aspx. Accessed June 5, 2019.
- Wade, M. R., & Tingling, P. (2005). A new twist on an old method: a guide to the applicability and use of web experiments in information systems research. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, 36(3), 69–88.
- Wakefield, M., Balch, G. I., Ruel, E., Terry-McElrath, Y., Szczypka, G., Flay, B., & Clegg-Smith, K. (2005). Youth responses to anti-smoking advertisements from tobacco-control agencies, tobacco companies, and pharmaceutical companies. *Journal of Applied Social Psychology*, 35(9), 1894–1910.
- Wired (2018). A bot panic hits Amazon Mechanical Turk. Available at: <https://www.wired.com/story/amazon-mechanical-turk-bot-panic/>. Accessed March 27, 2019.
- Witte, K., & Allen, M. (2000). A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health Education & Behavior*, 27(5), 591–615.
- Zephoria (2019). The top20 valuable Facebook statistics. Available at: <https://zephoria.com/top-15-valuable-facebook-statistics/>. Accessed March 29, 2019.
- Zikmund, W. G., Babin, B. J., Carr, J. C., & Griffin, M. (2013). *Business research methods*. Cengage Learning.

Davide C. Orazi is an Assistant Professor of Marketing at Monash Business School, Australia. His primary research focus is on consumer psychology, narrative theory, and protection motivation applied to information security, health communications, and social marketing. His research has appeared in *International Journal of Research in Marketing*, *Journal of Business Research*, *Journal of Business Ethics*, *European Journal of Marketing*, and *Journal of Advertising Research*, among many others. He serves on the Editorial Review Board of the *European Journal of Marketing*.

Allen C. Johnston is an Associate Professor of Management Information Systems in the Department of Information Systems, Statistics, and Management Science within the Culverhouse College of Commerce at the University of Alabama. The primary focus of his research is in the areas of behavioral information security, privacy, data loss prevention, collective security, and innovation. His research can be found in such outlets as *MIS Quarterly*, *Journal of the AIS*, *European Journal of Information Systems*, *Information Systems Journal*, *Communications of the ACM*, *Journal of Organizational and End User Computing*, *Information Technology and People*, and *DATABASE for Advances in Information Systems*. He currently serves as AE for *European Journal of Information Systems and Decision Sciences Journal*, as well as serving on the Editorial Review Board for *DATABASE for Advances in Information Systems*. He is a founding member and current Vice Chair of the IFIP Working Group on Information Systems Security Research.